exmjaxmbi

August 21, 2024

1 NaiveBayes Homework - Answer

Aug 20, 2024

1.1 1. Bài toán

Cho tập dữ liệu về bệnh ung thư: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Trong đó là các thông tin về đặc điểm của tế bào đã được ghi nhận thành các thuộc tính (radius_mean, texture_mean, ...), thể hiện dưới dạng bảng. Cùng với đó là các chẩn đoán (diagnosis) liệu tế bào đó có phải là tế bào ung thư hay không.

Hãy xây dựng mô hình Naive Bayes để dự đoán liệu một tế bào với các đặc điểm cho trước có phải là một tế bào ung thư hay không?

1.2 2. Import các thư viện cần thiết và load dữ liệu từ file

[ ]:
```
# Import các thư viện cần thiết
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

[ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[ ]:
```
%cd /content/drive/MyDrive/Code_VinBigData_2024/data
```

```
/content/drive/MyDrive/Code_VinBigData_2024/data
```

[ ]:
```
# Load data
data = pd.read_csv("cancer_data.csv")
data.head()  # Quan sát dữ liệu cho thấy, tập dữ liệu đã được xử lí khá "sạch"
```

[ ]:
```
data.info()
```

[ ]:
```
# Loại trường không cần thiết: id, unnamed
data = data.drop(["id", "Unnamed: 32"], axis=1)
```

[ ]:
```
# Trực quan hóa về dữ liệu: u ác tính và u lành:
M = data[data.diagnosis == "M"]  # dữ liệu ứng với u ác tính
B = data[data.diagnosis == "B"]  # dữ liệu ứng với u lành

plt.title("Trực quan dữ liệu u ác tính và u lành")
plt.xlabel("Radius Mean")
plt.ylabel("Texture Mean")
plt.scatter(M.radius_mean, M.texture_mean, color="red", label="U ác", alpha=0.3)
plt.scatter(B.radius_mean, B.texture_mean, color="lime", label="U lành",␣
 ↪alpha=0.3)
plt.legend()
plt.show()
```

1.3 Tiền xử lí trước khi huấn luyện mô hình:

[ ]:
```
# Câu hỏi: Đưa trường chẩn đoán từ dạng chữ thành dạng số
# Code ############

##########################################################
```

[ ]:
```
# Câu hỏi: Tách thuộc tính và nhãn:

# Code ################

##################
```

[ ]:
```
# Chuẩn hóa các trường của X
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
sc.fit(X)
```

[ ]:
```
# Câu hỏi: Chia dữ liệu train/test:
from sklearn.model_selection import train_test_split

# Code ###################

##############################
```

1.4 Xây dựng và đánh giá mô hình Naive Bayes:

[ ]:
```
# Câu hỏi: Multinormial NB:

# Code ###################################

########################################
```

[ ]:
```
# Câu hỏi: Gaussian NB:

# Code ########################################

##############################################
```

[ ]:
```
# Câu hỏi: Dùng GNB để dự đoán thử một sample = phần tử đầu tiên của x_test:

# Code ########################################3

################################################
```